Powered by Futurense

AIL1020 Foundations of Statistics & Probability

Module 02

How to interpret correlation

Dr. Rajlaxmi Chouhan
Associate Professor, Department of Electrical Engineering
IIT Jodhpur

Learning Outcomes

Select and justify the appropriate correlation method

Interpret correlation results based on the case and variables

Value of correlation

The strength of a correlation is based on how close the value of ***r*** is to ±1:

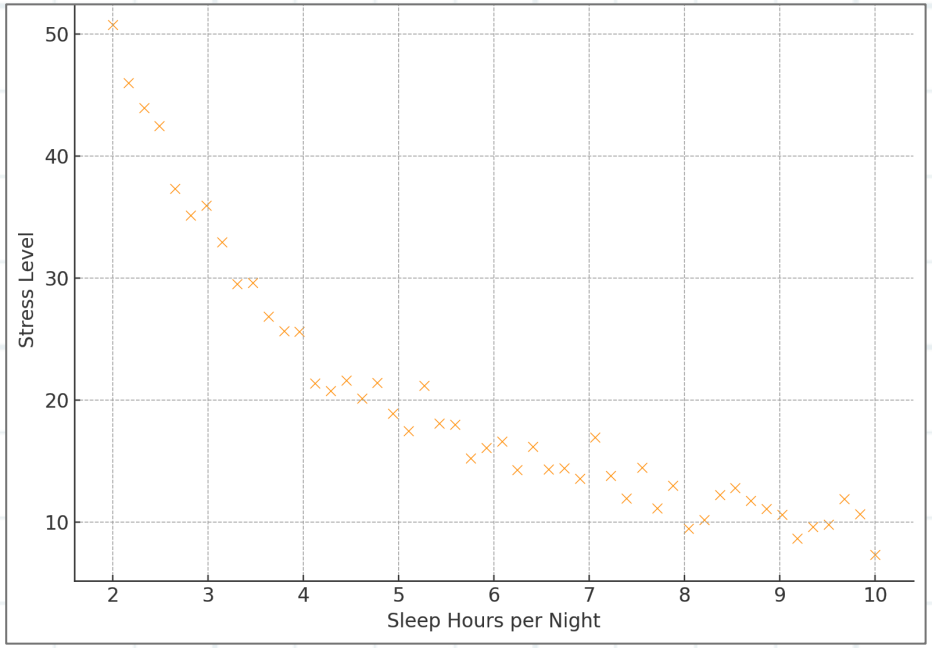
r-value Range	Interpretation
0.00 to ±0.10	No or negligible
±0.10 to ±0.30	Weak
±0.30 to ±0.50	Moderate
±0.50 to ±0.70	Strong
±0.70 to ±1.00	Very Strong

Choosing the right type of correlation is important!

1

When Pearson Gives Misleading Results

Scenario: A professor analyzes the relationship between *stress levels* and *sleep hours* among students.



*Data is **non-linear** but **monotonic.***

Stress Level:
1 to 10 scale
(Ordinal)

Sleep hours:
1 to 10 hours
(Continuous)

Choosing the right type of correlation is important!

Powered by Futurense

1

When Pearson Gives Misleading Results

Scenario: A professor analyzes the relationship between *stress levels* and *sleep hours* among students.

*Data is **non-linear** but **monotonic.***

(as sleep decreases, stress increases, but not at a constant rate)

Pearson's *r*: -0.90
suggests a strong linear relationship (misleading!)

Spearman *r* = -0.97
even stronger, capturing the monotonic trend

Stress Level:
1 to 10 scale
(Ordinal)

Sleep hours:
1 to 10 hours
(Continuous)

When are these correlation coefficients equivalent?

Powered by Futurense

2

All correlation methods (Pearson, Spearman, Kendall) tend to **converge** when data is,

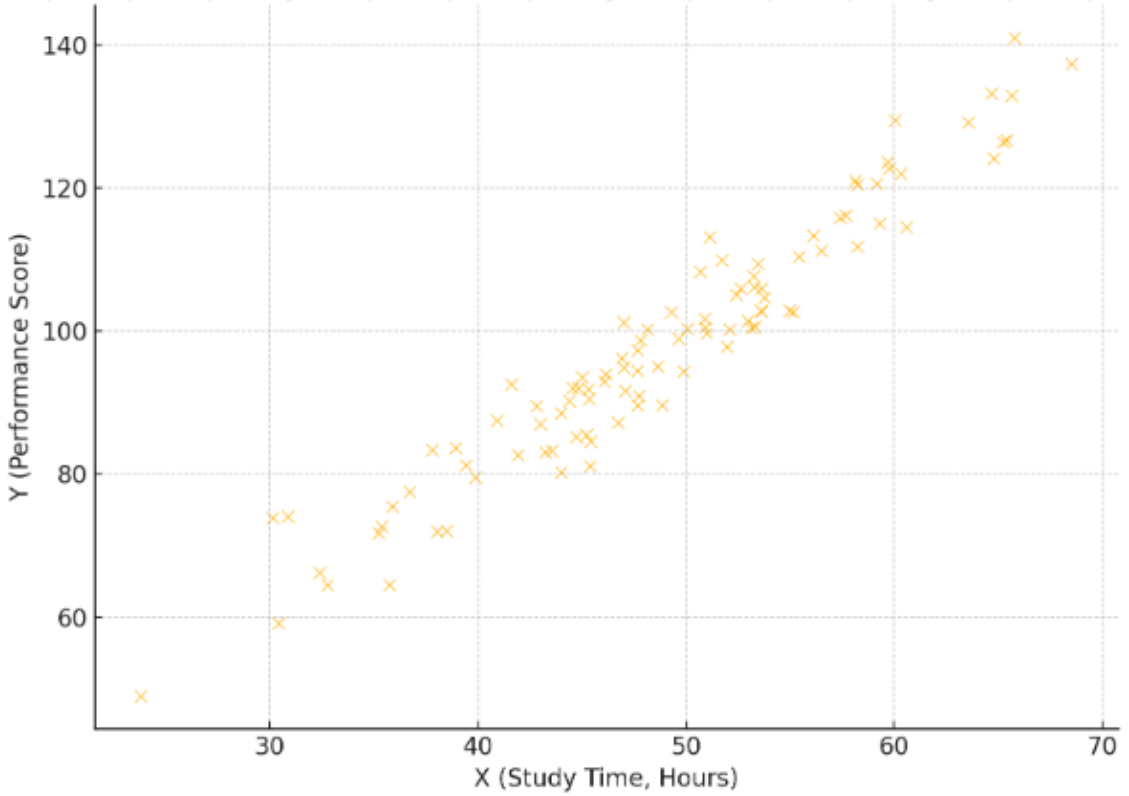
Continuous

Linearly related

Free from significant outliers

Monotonic and normally distributed

especially with large sample sizes



When binary data pretends to be continuous

3 Let's say a psychology researcher wants to explore whether **gender** influences **motivation**.

Data

Motivation scores (1 – 5) *Gender* (Code: 0 for male, 1 for female)

If the researcher finds Pearson's Correlation, and r = -0.268.

If the researcher finds Point Biserial Correlation, and r_{pb} = -0.268.

The values are identical in this case because point-biserial is a special case of Pearson's *r*.

⚠️ **Interpretation using Pearson**

Weak to Moderate negative correlation → "as gender increases, motivation decreases"

Misleading! Gender variable here is binary has **no inherent numerical order or distance!**

Powered by Futurense

When binary data pretends to be continuous

3 Let's say a psychology researcher wants to explore whether **gender** influences **motivation**.

Data

Motivation scores (1 – 5) *Gender* (Code: 0 for male, 1 for female)

If the researcher finds Pearson's Correlation, and $r = -0.268$.

If the researcher finds Point Biserial Correlation, and $r_{pb} = -0.268$.

✅ **Interpretation using Point Biserial**

In this dataset, there appears to be a **weak-to-moderate negative** correlation between being female and motivation score.

When binary data pretends to be continuous

3

Let’s say a psychology researcher wants to explore whether **gender** influences **motivation**.

Point-biserial **is Pearson in disguise** for binary vs. continuous — but using the correct method ensures you understand what kind of relationship you're analyzing.

What’s the role of chance here? Is this result statistically significant?

Powered by Futurense

Phi Coefficient and the danger of unbalanced data

4 Evaluation of a ***mentorship program***

100 students
- 95 joined → 85 passed the final exam
- 05 did not join → 03 passed the final exam

Since both variables are binary: **program participation** and **pass/fail**

Phi coefficient $\phi = 0.05 \rightarrow$ Suggests **very weak correlation**

"Mentorship has no effect." ⚠️

Phi Coefficient and the danger of unbalanced data

4

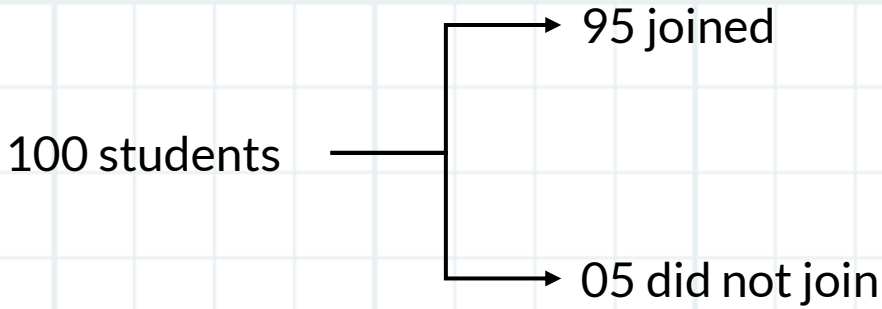
Evaluation of a ***mentorship program***

Since both variables are binary: **program participation** and **pass/fail**

Phi coefficient $\phi = 0.05 \rightarrow$ Suggests **very weak correlation**

A deeper look reveals:

- The dataset is **highly imbalanced** (only 5 students in the "No Program" group)
- This skews the **contingency table**, reducing the power of the Phi coefficient.



A better approach? Use a **chi-square test** or **odds ratio.**
Find out more about these approaches.

ϕ coefficient can **mask real effects** when group sizes are unequal. (Always check the distribution)

Is finding just correlation coefficient enough for interpretation?

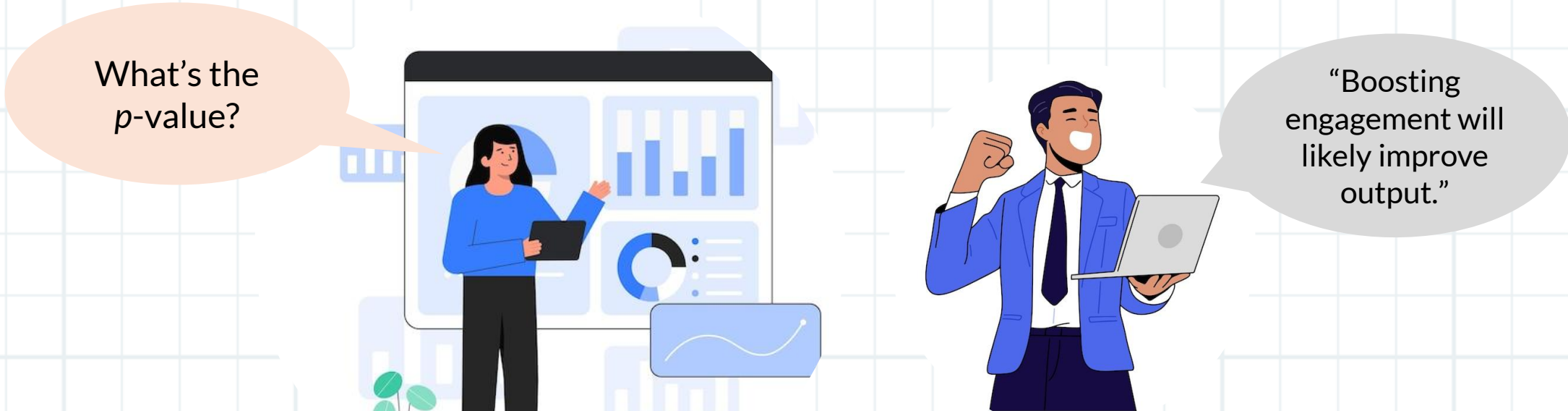
No

Scenario:
A mid-level manager at a tech company notices that teams with higher ***employee engagement*** seem to get better ***performance ratings***.

Quick survey across **8 team leads**

- Engagement (on a 10-point scale)
- Team performance (targets, etc.)

Pearson correlation r = 0.35!!



Is finding just correlation coefficient enough for interpretation?

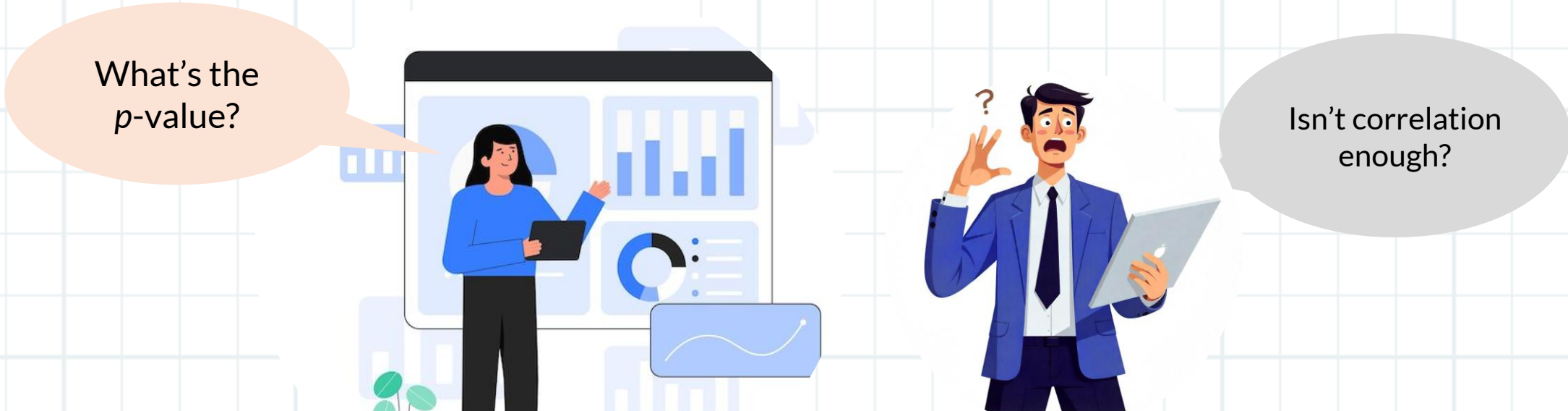
No

Scenario:
A mid-level manager at a tech company notices that teams with higher ***employee engagement*** seem to get better ***performance ratings***.

Quick survey across **8 team leads**

- Engagement (on a 10-point scale)
- Team performance (targets, etc.)

Pearson correlation r = 0.35!!



Is finding just correlation coefficient enough for interpretation?

No

Scenario:
A mid-level manager at a tech company notices that teams with higher ***employee engagement*** seem to get better ***performance ratings***.

Quick survey across **8 team leads**

- Engagement (on a 10-point scale)
- Team performance (targets, etc.)

A correlation coefficient tells you the **strength of the relationship** — but it doesn't tell you if the relationship is **statistically significant**.

With only 8 data points, even a moderate correlation might be due to **random chance**.

Pearson correlation r = 0.35!!

Even though r = 0.35 suggests a relationship, **we can't rule out randomness**.

p-value = 0.32

Find ***r*** AND ***p*-value**

➔ p-value shows the probability that this correlation came just by chance.
➔ If $p > 0.05$, No 'statistical significance'

Powered by Futurense

Summary

Do not treat a binary variable like a number just because it's coded 0/1.

Use the right Correlation Method depending on the data type.

Always look at the sample size, distribution, and role of probability before interpreting results.

Always match your correlation method to your data types — and understand what the statistic is actually telling you.

Powered by

Futurense

End of Module 02

Next Module: ***Probability***